Alcohol Language Corpus: a publicly available large corpus of alcoholized speech

F. Schiel[1], Chr. Heinrich[1], S. Barfüßer[1], Th. Gilg[2]

[1]*Institut für Phonetik und Sprachverarbeitung*

[2]*Institut für Rechtsmedizin*

Ludwig-Maximilians-Universität, München, Germany

{schiel|heinrich|bine}@phonetik.uni-muenchen.de

thomas.gilg@med.uni-muenchen.de

The Alcohol Language Corpus (ALC) contains German spoken language recorded from a large number of female and male speakers being sober and intoxicated. It was produced by the Bavarian Archive of Speech Signals (BAS) in the years 2007-2010 and is publicly available for research and educational purposes (see http://www.bas.uni-muenchen.de/Bas).

Although a number of studies have been conducted dealing with speech under the influence of alcohol (see (Chin & Pisoni, 1997) for an overview) there exists no publicly available reference database on which researchers might replicate published results and conduct their own investigations. The ALC project aims to provide such a corpus with the following characteristics:

- balanced recordings of male and female speakers
- minimum of 60 speakers per gender to allow reliable F-statistics on features where only one measurement per speaker is possible
- three different speech styles: read, spontaneous and elicited command&control
- coverage of age between 21 to 65
- coverage of blood alcohol concentration (BAC) from 0.03 to 0.17%
- linguistic and phonetic annotation and segmentation, para-linguistic tagging (e.g. disfluencies)
- consistent acoustic environment and dialog partners accross different conditions

Another motivation for ALC is the hypothesis that alcoholic intoxication may automatically be detected by a standard speech interface within an automobile. The basic idea is that the system observes the normal (sober) voice of the driver during his/her daily interactions, detects characteristic feature deviations that might indicate alcoholic intoxication (Schiel & Heinrich, 2009) and then acts preemptively to protect the driver and other traffic members. Consequently ALC has been recorded in the automotive environment and contains 30% typical command&control sequences of in-car speech interfaces.

In its current distribution (version 2.1) ALC contains 14965 speech recordings from 70 female and 68 male speakers (final estimate is 162). Each speaker is tested for breath alcohol and blood alcohol concentration. Each speaker has uttered roughly 6min of speech intoxicated and 12min in sober state (30 and 60 recording items). The content covers digit strings (phone/credit card numbers), license plates, addresses, tongue twisters, picture descriptions, interview style question answering, free dialogues and command&control (both read and semi-spontaneous). The vocabulary size is 14013. Aside from the intoxication we register age, sex, height, weight, smoking habits, drinking habits, dialect origin, education, weather, acoustical environment and delay between recordings. The speech signal is captured by two microphones, a headset and a typical car built-in far field microphone (mouse micro). The total size of the corpus is 30GBytes.

In our contribution we will give details about the recruiting and recording process, the linguistic and phonetic annotation integrated into an Emu database, some interesting numbers derived from the corpus and the para-linguistic tagging and how to obtain the corpus from BAS.

References

Chin, S.B. and Pisoni, D.B. (1997). Alcohol and Speech. Academic Press.

Schiel, F. and Heinrich, Chr. (2009). Laying the Foundation for In-car Alcohol Detection by Speech. Proc. of the INTERSPEECH 2009, Brighton, UK, pp. 983-986.